First Elements on Knowledge Discovery guided by Domain Knowledge (KDDK)

Jean Lieber, Amedeo Napoli, Laszlo Szathmary, Yannick Toussaint

First Elements on Knowledge Discovery Guided by Domain Knowledge (KDDK)

Jean Lieber, Amedeo Napoli, Laszlo Szathmary, and Yannick Toussaint

LORIA (CNRS – INRIA – Universités de Nancy)
Équipe Orpailleur, Bâtiment B, BP 239
F-54506 Vandœuvre-lès-Nancy cedex, France
{''FirstName''.''LastName''}@loria.fr

Abstract. In this paper, we present research trends carried out in the Orpailleur team at LORIA, showing how knowledge discovery and knowledge processing may be combined. The knowledge discovery in databases process (KDD) consists in processing a huge volume of data for extracting significant and reusable knowledge units. From a knowledge representation perspective, the KDD process may take advantage of domain knowledge embedded in ontologies relative to the domain of data, leading to the notion of "knowledge discovery guided by domain knowledge" or KDDK. The KDDK process is based on the classification process (and its multiple forms), e.g. for modeling, representing, reasoning, and discovering. Some applications are detailed, showing how KDDK can be instantiated in an application domain. Finally, an architecture of an integrated KDDK system is proposed and discussed.

1 Introduction

In this presentation, we present research trends carried out within in the Orpailleur team at LORIA, showing multiple aspects of knowledge discovery and knowledge processing. The knowledge discovery in databases process –hereafter KDD– consists in processing a huge volume of data in order to extract knowledge units that are significant and reusable. Assimilating knowledge units to gold nuggets, and databases to lands or rivers to be explored, the KDD process can be likened to the process of searching for gold (in the same way, KDD is compared with archeology in [7]). This explains the name of the research team: the "orpailleur" denotes in French a person who is searching for gold in rivers or mountains. Moreover, the KDD process is iterative, interactive, and generally controlled by an expert of the data domain, called the *analyst*. The analyst selects and interprets a subset of the extracted units for obtaining knowledge units having a certain plausibility. As a person searching for gold and having a certain knowledge of the task and of the location, the analyst may use its own knowledge but also knowledge on the domain of data for improving the KDD process. Indeed, the objective of this paper is to show the role that can be played by domain knowledge within the KDD process.

From an operational point of view, the KDD process is performed within a KDD system including databases, data mining modules, and interfaces for interactions,

S. Ben Yahia et al. (Eds.): CLA 2006, LNAI 4923, pp. 22–41, 2008.

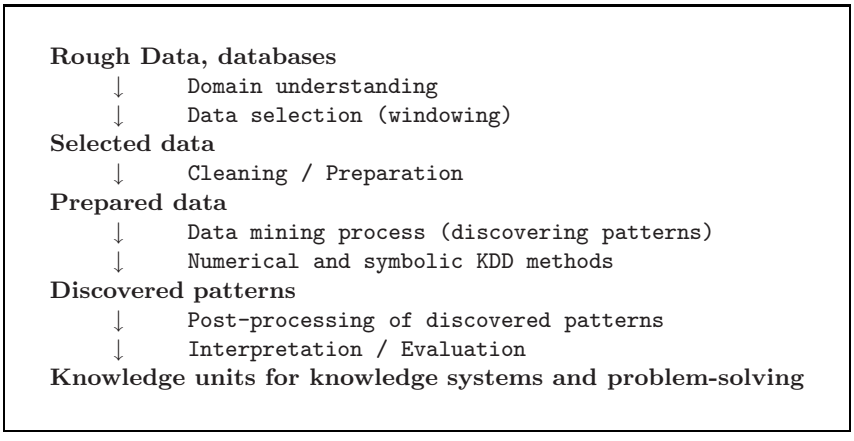

Fig. 1. From data to knowledge units: the objective of the knowledge discovery process is to select, prepare and extract knowledge units from different data sources. For effective reuse, the extracted knowledge units have to be represented within an adequate knowledge representation formalism.

e.g. editing and visualization. The KDD process is based on three main operations: selection and preparation of the data, data mining, and finally interpretation of the extracted units (see Figure 1).

A way for the KDD process to take advantage of domain knowledge is to be in connection with an *ontology* relative to the domain of data, a step towards the notion of *knowledge discovery guided by domain knowledge* or KDDK. In the KDDK process, knowledge units that are extracted have still a life after the interpretation step: they must be represented in an adequate knowledge representation formalism for being integrated within an ontology and reused for problem-solving needs. In this way, the results of the knowledge discovery process may be reused for enlarging existing ontologies. The KDDK process shows that knowledge representation and knowledge discovery are two complementary tasks: *no effective knowledge discovery without domain knowledge!*

Hereafter, we present various instantiations of the KDDK process that are all based on the idea of *classification*. Classification is a polymorphic process involved in various tasks, e.g. modeling, mining, representing, and reasoning (see also [42,10,53]). Accordingly, a knowledge-based system may be designed, fed up by the KDDK process, and used for problem-solving in application domains, e.g. agronomy, astronomy, biology, chemistry, and medicine (these application domains are studied in the Orpailleur team). A special mention has to be made for Semantic Web activities, involving in particular text mining, content-based document mining, and intelligent information retrieval (see for example [16,8,41]).

The paper is organized as follows. In the next section, symbolic methods for KDD and the CORON platform are introduced. Then, research trends in KDDK are presented and detailed, showing how knowledge can be embedded at each step of the KDD process. In the last section, an architecture for an integrated

KDDK system is described, and the KDD and KDDK processes are studied with respect to this integrated architecture.

2 Methods and Systems for KDD

The KDD process is based on *data mining methods* that are either symbolic or numerical [19,20,14]. The methods that are used in the Orpailleur team are the following (mainly symbolic methods):

- Symbolic methods based on lattice-based classification (concept lattice design or formal concept analysis [18]), frequent itemsets search, and association rule extraction [35]. These symbolic methods are more deeply described in the next subsection.
- Numerical methods based on second-order Hidden Markov Models (HMM2, initially designed for pattern recognition) [30,29]. Hidden Markov Models have good capabilities for locating stationary segments, and are mainly used for mining temporal and spatial data. The CarottAge system[1] is developed in the Orpailleur team for analyzing numerical spatio-temporal data.

In the following, the focus is on symbolic KDD methods. However, an ongoing research work holds on the combination of symbolic and numerical methods, that is discussed in section 3.4.

2.1 Lattice Design, Itemset Search and Association Rule Extraction

Classification problems can be formalized by means of a class of individuals (or objects), a class of properties (or attributes), and a binary correspondence between the two classes, indicating for each individual-property pair whether the property applies to the individual or not [3,18,8]. The properties may be features that are present or absent, or the values of a property that have been transformed into binary variables. Lattice-based classification relies on the analysis of such binary tables and may be considered as a symbolic data mining technique to be used for extracting (from a binary database) a set of concepts organized within a hierarchy (i.e. a partial ordering). The extraction of frequent itemsets, i.e. sets of properties or features of data occurring together with a certain frequency, and of association rules emphasizing correlations between sets of properties with a given confidence, are related activities.

The search for frequent itemsets and association rule extraction are well-known symbolic data mining methods. These processes usually produce a large number of items and rules, leading to the associated problems of "mining the sets of extracted items and rules". Some subsets of itemsets, e.g. frequent closed itemsets (FCIs), allow to find interesting subsets of association rules, e.g. informative association rules. This is why several algorithms are needed for mining data depending on specific applications.

[1] CarottAge is a free software developed in the Orpailleur team, with a GPL license since 2002, see `http://www.loria.fr/~jfmari/App/`

2.2 Rare Itemsets and Rules

Among useful patterns extracted from a database, frequent itemsets are usually thought to unfold "regularities" in the data, i.e. they are the witnesses of recurrent phenomena and they are consistent with the expectations of the domain experts. In some situations however, it may be interesting to search for "rare" itemsets, i.e. itemsets that do not occur frequently in the data (contrasting frequent itemsets). These correspond to unexpected phenomena, possibly contradicting beliefs in the domain. In this way, rare itemsets are related to "exceptions" and thus may convey information of high interest for experts in domains such as biology or medicine. For example, suppose an expert in biology is interested in identifying the cause of cardiovascular diseases (CVD) for a given database of medical records. A frequent itemset such as "{elevated cholesterol level, CVD}" may validate the hypothesis that these two items are frequently associated, leading to the possible interpretation "people having a high cholesterol level are at high risk for CVD". On the other hand, the fact that "{vegetarian, CVD}" is a rare itemset may justify that the association of these two itemsets is rather exceptional, leading to the possible interpretation "vegetarian people are at a low risk for CVD". Moreover, the itemsets {vegetarian} and {CVD} can be both frequent, while the itemset {vegetarian, CVD} is rare.

Rare cases deserve special attention because they represent significant difficulties for data mining algorithms. The underlying mining problems have been studied in detail, with different names, e.g. exceptions, negative rules (see for example [28,43,52,54,37]. These approaches are, most of the time, based on adaptions of the general levelwise Apriori algorithm. These methods typically retrieve large sets of rare itemsets and association rules, but these methods may remain incomplete –rare associations are be discovered– either due to an excessive computational cost or to overly restrictive definitions. Thus, such methods may fail to collect a large number of potentially interesting patterns.

By contrast, a framework is proposed in [45,44,48] is specifically dedicated to the extraction of rare itemsets. It is based on an intuitive yet formal definition of rare itemset. Its goal is to provide a theoretical foundation for rare pattern mining, with definitions of reduced representations and complexity results for mining tasks, as well as to develop an algorithmic tool suite (within the Coron platform, see next subsection) together with the guidelines for its use. The method, for computing all rare itemsets is based on two main steps. The first step thereof is the identification of the *minimal rare itemsets* with an optimized method that limits the exploration to frequent generators only (minimal rare itemsets jointly act as a minimal generation seed for the entire rare itemset family). The second step is performed to restore all rare itemsets from minimal rare itemsets.

2.3 The Coron Platform

The Coron platform[2] is currently developed in the Orpailleur team [46,47]. The platform is composed of three main modules: (i) Coron-base, (ii) AssRuleX,

[2] `http://coron.loria.fr`

(iii) pre-processing and post-processing modules. The CORON-base module is aimed at extracting different kinds of itemsets, e.g. frequent itemsets, frequent closed itemsets, minimal generators, etc. The module contains a collection of important data mining algorithms, such as Apriori, Close, Pascal, Titanic, Charm, Eclat, together with adapted algorithms such as Zart and Eclat-Z (plus some others). This large collection of (efficient) algorithms is one of the main characteristics of the CORON platform. Knowing that each of the algorithms has advantages and disadvantages with respect to the form of the data to be mined, and since there is no universal algorithm for processing any arbitrary dataset, the CORON-base module offers to the user the choice of the algorithm that is the best suited for his needs.

The second module of the system, ASSRULEX (Association Rule eXtractor) generates different sets of association rules, such as informative rules, generic basis, and informative basis.

For supporting the whole life-cycle of a data mining task, the CORON platform proposes modules for cleaning the input dataset and reduce its size if necessary. The module RULEMINER facilitates the interpretation and the filtering of the extracted rules. The association rules can be filtered by (i) attribute, (ii) support, and/or (iii) confidence.

The CORON platform is developed entirely in Java, allowing portability. The system is operational, and has been tested within several research projects within the team [12,31].

2.4 A Data-Mining Methodology with the Coron Platform

A methodology was initially designed for mining biological cohorts, but it can be generalized to any kind of database. It is worth to mention that the whole KDDK process is guided by an analyst. The role of the analyst is important with respect to the following tasks: selecting the data and interpreting the extracted units. This methodology is associated to the CORON platform, that offers various tools necessary for its application in a single platform (nevertheless another platform can be used).

The methodology consists of the following steps: **(1)** Definition of the study framework, **(2)** Iterative step: data preparation and cleaning, pre-processing step, processing step, post-processing step, validation of the results and Generation of new research hypotheses, feedback on the experiment. The life-cycle of the methodology is shown in Figure 2.

Definition of the Study Framework.
The analyst defines a specific field for the analysis (called hereafter "framework"). Thus, he may choose the type of data he wants to work on, e.g. biological data, genetic data, or both, unrelated individuals or families, focus on a special metabolic network or on a particular syndrome.

Iterative Step.
Data preparation and cleaning. Data cleaning is necessary. This step includes the detection and the possible removal of incomplete and out-of-range

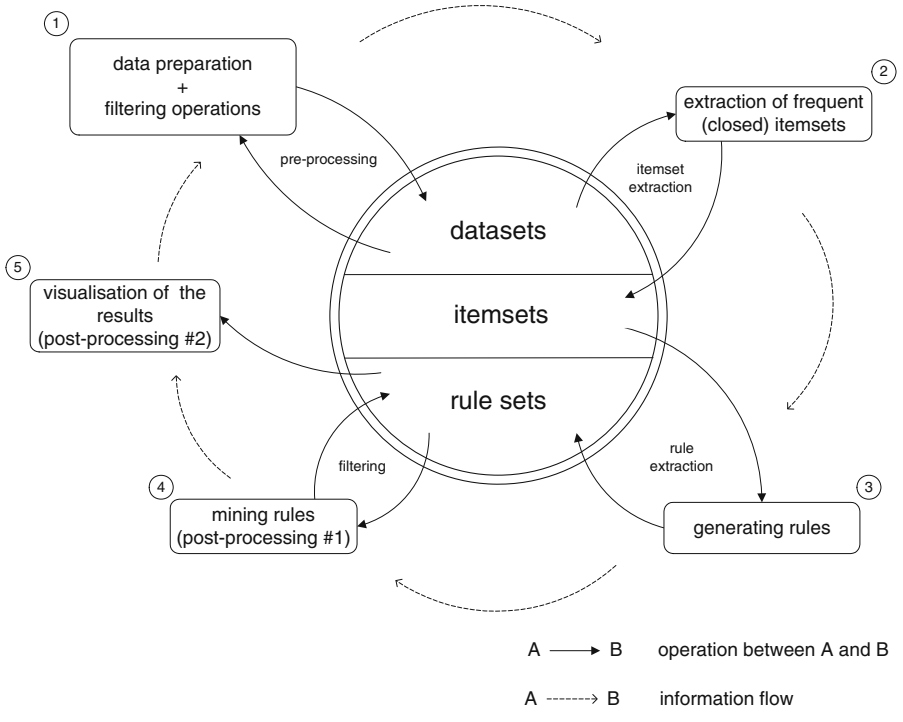

Fig. 2. The life cycle of KDD within Coron

values. Moreover, several actions for converting the data can be done at this step, such as:

(1) *Addition/creation* of new attributes for helping the extraction of association rules by combining attributes (intersection, union and complementary).

(2) *Deletion* of attributes that are not interesting in the chosen biological framework. This option is close to the projections described below.

(3) *Discretization* for transforming continuous data into Boolean values, e.g. by using a threshold defined in the literature, or by separating values of each continuous variable into quartiles.

Data filtering (pre-processing). Several actions can be carried out that correspond to operations in set theory: complement, union and intersection (with operations of additions and projections).

(1) *Apply projections*:

on the rows: i.e. selecting individuals with one or more attributes specified by the expert,

on the columns: i.e. selecting (or deleting) some attributes.

(2) *Consider the complement* of a set of individuals satisfying a rule, defined by the set of individuals that do not satisfy this rule.

The output of the filtering process is considered as a new dataset on which data mining procedures can be applied again.

Applying the data mining procedure. This methodology is related to symbolic data mining methods, as, in particular, frequent itemset search and association rule extraction. With the help of the analyst, the necessary thresholds values can be set for quality measures such as the minimum support and the minimum confidence for generating frequent itemsets and association rules, respectively. As the process is iterative and interactive, the analyst can change these thresholds during a next iteration to carry out new experiments.

Post-processing. After filtering and visualizing the rules, those rules containing the most interesting attributes can be found. If a less relevant attribute is always present in the rules, it can be considered as "noisy", and removed from the input dataset. This means that the dataset is another time modified for a new association rule extraction.

The iterative step can be repeated until the most relevant rules are found. The interpretation of the analyst is mobilized both for rule mining and result visualization.

Rule-mining. In the rule mining step, the analyst has also to make several choices:

- *Choosing rules with a specific form*: e.g. selecting rules that only have one attribute on their left side.
- *Selecting rules with an attribute of interest* from the point of view of the analyst, on the left hand side, on the right hand side, or on both sides.
- *Classifying the extracted rules* in ascending or descending order according to their support or confidence values, or according to other statistical values [9].
  The classification of rules mining step may be dependent on numerical measures, e.g. support and confidence, or on domain knowledge as shown in some experiments [21].
- *Selecting rules* with a support belonging to a given interval $[a, b]$; returning rules with a support less than (or more than) or equal to a given value c. These selections can also be applied with the other statistical measures cited above.

Visualization of the results. A visualization method adapted to symbolic data mining method procedure has to be chosen. For frequent itemset search leading to the extraction of less frequent itemsets, concept lattices may be used beneficially [22].

Validation of the results and generation of new research hypotheses. The evaluation of the rules can be done either by statistical tests, data analysis methods, i.e. automatic classification, component analysis, or with knowledge-based methods, e.g. classification-based reasoning, formal concept analysis. The generated results allow the expert to suggest new directions of research. Accordingly, these new hypotheses are tested by new experiments, for example,

managed at the biological level, like genetic epidemiological studies or wet laboratory experiments.

3 Research Directions for KDDK

The principle summarizing KDDK can be read as follows: going "from complex data units to complex knowledge units guided by domain knowledge" (KDDK) or "knowledge with/for knowledge". This principle is discussed below, along research activities such as graph mining, spatio-temporal data mining, text mining and Semantic Web, knowledge discovery in life sciences, combining symbolic and numerical data mining methods for hybrid mining, and finally mining a knowledge base, a kind of "meta-knowledge discovery process". All these research activities share the fact that the mining process is guided and enhanced by domain knowledge (similar ideas are also discussed in [11,53]).

3.1 KDDK and the Mining of Complex Data

Lattice-based classification, formal concept analysis, itemset search and association rule extraction, are suitable paradigms for symbolic KDDK, that may be used for real-sized applications [51]. Global improvements may be carried on the ease of using of the data mining methods, on the efficiency of the methods [24], and on adaptability, i.e. the ability to fit evolving situations with respect to the constraints that may be associated with the KDDK process. Accordingly, the research work presented hereafter is in concern with the extension of symbolic methods to complex data, e.g. objects with multi-valued attributes, relations, graphs, texts, and real world data.

KDDK in databases of chemical reactions.
The mining of chemical chemical reaction databases is an important task for at least two reasons (see also [23]): (i) the first reason is the challenge represented by this task regarding KDDK to be set on, (ii) the second reason lies in the industrial needs that can be met whenever substantial results are obtained. Chemical reactions are complex data, that may be modeled as undirected labeled graphs. They are the main elements on which synthesis in organic chemistry relies, knowing that synthesis —and accordingly chemical reaction databases— are of first importance in chemistry, but also in biology, drug design, and pharmacology. From a problem-solving point of view, synthesis in organic chemistry must be considered at two main levels of abstraction: a strategic level where general synthesis methods are involved –a kind of meta-knowledge– and a tactic level where specific chemical reactions are applied. An objective for improving computer-based synthesis in organic chemistry is aimed at discovering general synthesis methods from currently available chemical reaction databases for designing generic and reusable synthesis plans.

A preliminary research work has been carried on in the Orpailleur team [5], based on frequent levelwise itemset search and association rule extraction, and

applied to standard chemical reaction databases. This work has given substantial results for the expert chemists. At the moment, for extending this first work, a graph-mining process is used for extracting knowledge from chemical reaction databases, directly from the molecular structures and the reactions themselves, This research work is currently under development, in collaboration with chemists, and is in accordance with needs of chemical industry [38].

KDDK and the mining of spatio-temporal data.
Temporal and spatial data are complex data to be mined because of their internal structure, that can be considered as multi-dimensional. Indeed, spatial data may involve two or three dimensions for determining a region and complex relations as well for describing the relative positions of regions between each others (as in the RCC-8 theory for example [26,36]). Temporal data may present a linear but also a two-dimensional aspect, when time intervals are taken into account and have to be analyzed (using Allen relations for example). In this way, mining temporal or spatial data are tasks related to KDDK. Spatial and temporal data may be analyzed with numerical methods such as Hidden Markov Models, but also with symbolic methods, such as levelwise search for frequent sequential or spatial patterns.

In the medical domain, the study of chronic diseases is a good example of KDDK process on spatio-temporal data. An experiment for characterizing the patient pathway using the extraction of frequent patterns, sequential and not sequential, from the data of the PMSI[3] system associated with the "Lorraine Region" is currently under investigation. Details on this work are given in [22].

3.2 KDDK, Text Mining and Semantic Web

KDDK and text mining.
The objective of a text mining process is to extract new and useful knowledge units in a large set of texts [21,10,9]. The text mining process shows some specific characteristics due to the fact that texts are complex objects written in natural language. The information in a text is expressed in an informal way, following linguistic rules, making the mining process more complex. To avoid information dispersion, a text mining process has to take into account –as much as possible– paraphrases, ambiguities, specialized vocabulary, and terminology. This is why the preparation of texts for text mining is usually dependent on linguistic resources and methods. In addition, from a KDDK perspective, the text mining process is aimed at extracting new knowledge units from texts with the help of background knowledge. The interpretation of a text relies on knowledge units shared by the authors and the readers. A part of these knowledge units is expressed in the texts and may be extracted by the text mining process. Another part of these knowledge units, background knowledge, is not explicitly expressed in the text and is useful to relate notions present in a text, to guide and to help

[3] For "Programme de Médicalisation des Systèmes d'Informations". This is the name of the information system collecting the administrative data for an hospital.

the text mining process. Background knowledge is encoded in a knowledge base associated to the text mining process. Text mining is especially useful in the context of semantic Web, for manipulating textual documents by their content.

The studies on text mining carried out in the Orpailleur team hold on real-world texts in application domains such as astronomy, biology and medicine, using mainly symbolic data mining methods such as i.e. frequent itemset search and association rule extraction [4]. This is in contrast with text analysis approaches dealing with specific language phenomena. The language in texts is considered as a way for presenting and accessing information, and not as an object to be studied for its own. In this way, the text mining process may be involved in a loop used to enrich and to extend linguistic resources. In turn, linguistic and ontological resources can be exploited to guide a "knowledge-based text mining process".

KDDK within the context of Semantic Web.
Semantic Web constitutes a good platform for experimenting ideas on knowledge discovery –especially text mining–, knowledge representation and reasoning. In particular, the knowledge representation language associated with the Semantic Web is the OWL language, based on description logics (or DLs, see [2]). In OWL, knowledge units are represented within concepts (or classes), with attributes (properties of concepts, or relations, or roles), and individuals. The hierarchical organization of concepts (and relations) relies on a subsumption relation that is a partial ordering. The inference services are based on subsumption, concept and individual classification, two tasks related to "classification-based reasoning". Concept classification is used for inserting a new concept at the right location in the concept hierarchy, searching for its most specific subsumers and its most general subsumees. Individual classification is used for recognizing the concepts an individual may be an instance of. Furthermore, classification-based reasoning may be extended into case-based reasoning (CBR), that relies on three main operations: retrieval, adaptation, and memorization. Given a target problem, retrieval consists in searching for a source (memorized) problem similar to the target problem. Then, the solution of the source problem is adapted to fulfill the constraints attached to the target problem. When there is enough interest, the target problem and its solution may be memorized in the case base to be reused. In the context of a concept hierarchy, retrieval and adaptation may be both based on classification (and "adaptation-guided retrieval" [17]).

In the framework of Semantic Web, the mining of textual documents on the Web, or "Web document mining" [6], can be considered from two main points of view: (i) mining the content of documents, involving text mining, (ii) mining the internal and external –hypertext links– structure of pages, involving information extraction. Web document mining is a major technique for the semi-automatic design of real-scale ontologies, the backbone of Semantic Web. In turn, ontologies are used for annotating the documents, enhancing document retrieval and document mining. In this way, Web document mining improves annotation, retrieval, and the understandability of documents, with respect to their structure and their

content. The extracted knowledge units can then be used for completing domain ontologies, that, in turn, guide text mining, and so on.

A research carried on in the team aims at understanding the structure of documents for analyzing and for improving text mining. The design of a system for extracting information units –that have to be turned into knowledge units after interpretation– from Web pages involves a wrapper-based machine learning algorithm combined with a classification-based reasoning process, taking advantage of a domain ontology implemented within the Web Ontology Language (OWL). The elements returned by the process are used as "semantic annotations" for understanding and manipulating the documents with respect to their structure and content [50,49]. The application domain of this research work is the study of research themes in the European Research Community. This study supports the analysis of research themes and detection of research directions.

3.3 KDDK for Life Science: Organizing and Navigating Biological Sources

The application domains that are currently investigated at the moment by the Orpailleur team are related with life sciences, with a particular emphasis on biology (bioinformatics) and medicine. Indeed, there are various reasons explaining why life sciences are a major application domain. In general, life sciences are getting more and more importance as a domain application for computer scientists. In this context, the collaboration between biologists and computer scientists is very active, and the understanding of biological systems provides complex problems for computer scientists. When these problems are solved (at least in part), the solutions bring new ideas not only for biologists but also for computer scientists in their own research work. Thus, advances in research appear on both sides, life and computer sciences.

Knowledge discovery is gaining more and more interest and importance in life sciences for mining either homogeneous databases (DBs) such as protein sequences or structures, heterogeneous DBs for discovering interactions between genes and environment, or between genetic and phenotypic data, especially for public health and pharmacogenomics domains. The latter case appears to be one main challenge in knowledge discovery in biology and involves knowledge discovery from complex data and thus KDDK. The interactions between researchers in biology and researchers in computer science improve not only knowledge about systems in biology, but knowledge about computer science as well. Solving problems for biologists using KDDK methods may involve the design of specific modules that, in turn, leads to adaptations of the KDDK process, especially in the preparation of data and in the interpretation of the extracted units.

A research work carried on in the team is in concern with the search and the access to relevant biological sources (including biological DBs) satisfying a set of given constraints, expressed with respect to concepts lying in a domain ontology –as in the BioRegistry repository [40]. The sources may be described in terms of these concepts, yielding a formal context, from which a concept lattice can be built [32]. Given a specific query, a lattice-based information retrieval process is

set on. The classification of the query in the lattice returns a ranked list of relevant sources, according to the characteristics of the sources with respect to the characteristics of the query (see [33]). The next step is to generalize the approach, and to use a "fuzzy concept lattice" and "fuzzy formal concept analysis" (see for example [39]). Moreover, studies hold on complex question answering methods taking into account fuzzy concept lattices, nested queries (intersection, union, and complement), analogical queries, and composition of answers elements. These techniques are still under study.

Another challenge is to extract knowledge from heterogeneous DBs for understanding interactions between clinical, genetic and therapeutic data. For example, a given genotype, i.e. a set of selected gene versions, may explain adverse clinical reactions (e.g. hyperthermy, toxic reaction...) to a given therapeutic treatment. This requires first the integration of both genomic and clinical data into a data warehouse on which KDDK methods have to be applied. This research work is connected with Semantic Web purposes, and in particular with the following elements: (i) data preparation and extracted units interpretation based on domain ontologies, (ii) knowledge edition for building and enriching domain ontologies, (iii) knowledge management for access to knowledge units, querying and reasoning (for problem-solving).

3.4 Combining Symbolic and Numerical Methods for KDDK

Why combining symbolic and numerical methods.
HMM2 have proved to be a valuable tool for extracting knowledge from complex numerical data, e.g. spatio-temporal data. In this way, the CarottAge system has been involved for data mining purposes in two main application domains, namely biology and agronomy. In collaboration with biologists, genome segmentation and interpretation have been investigated [15]. In collaboration with agronomists, spatial and temporal land-use data have been mined for extracting and understanding crop successions, i.e. the way how crops are carried out during a given period of time [25,30]. In these two applications, the effort has focused on two main points, with respect to the questions of the biologists and of the agronomists: (i) the elaboration of a mining process for extracting dependencies in temporal and spatial data involving an unsupervised classification process based on HMM2, (ii) the specification of associated and adequate visualization tools giving a synthetic view of the extraction process results to the experts in charge of interpreting the extracted classes and/or of specifying new experiment directions.

However, some operations remain very difficult to be carried out and could be eased using symbolic methods: (i) the modeling of the HMM2 process for a set of given data, (ii) the interpretation of units extracted by HMM2, (iii) the organization and the visualization of the extracted units for further reuse, e.g. as knowledge units in a knowledge-based system. A proposition is to combine HMM2 with symbolic methods, such as case-based reasoning and concept lattices, for helping the modeling and interpretation process.

A challenge is to set on a methodology for hybrid KDDK, coupling HMM2 and symbolic methods, that can be adapted and reused as a general KDDK method on various data, leading to a multi-functional and multi-purpose KDDK system.

Combining CBR and HMM.
Case-based reasoning seems to be especially interesting since researchers in an application domain often use their own knowledge or knowledge resulting from first experiments to improve steps within the data mining process, e.g. modeling and interpretation. In this way, the elements of the cases within the case-based reasoner can be composed of knowledge units about parameters of the HMM2, and as well of knowledge units on the design –modeling, data preparation–, and the interpretation –relying on ontological knowledge– of the HMM2. In addition, CBR can be of great interest for recording mining strategies that can be adapted and reused in similar situations. Indeed, a study on CBR for guiding mining scenarios in a given situation –with retrieval and adaptation of a similar situation– has not yet been carried on and should give substantial results. More generally, HMM2-based data mining process may take advantage of being coupled with CBR, that can be used at a strategic level for guiding the HMM2-based data mining process.

Combining concept lattices and HMM.
For their part, concept lattices can be used to organize and to visualize the results of the HMM2-based data mining process. The objects resulting of the application of the HMM2 process can be characterized by a set of properties. For example, in a spatio-temporal framework, space regions may be considered as objects and characteristics of the region at a given time can be considered as properties, yielding a kind of formal context. In addition, itemsets and association rules may also be extracted from such a context, offering an easy way of interpreting results of the HMM2 process.

The analysis of complex data in biology also calls for the coupling of symbolic and numerical data mining methods. There are complex data on which HMM2 show a good behavior, for recognizing and extracting regular structures. Such complex data hold on interactions between processes or agents, such as data from transcriptomic biochips –DNA chips or microarrays– experiments (used for extracting knowledge on interactions between plants and microorganisms). Still, an important objective of this kind of study is to investigate and to understand more deeply the modeling of biological systems, at symbolic and numerical levels.

3.5 Meta-knowledge Discovery of Mining Knowledge Bases

The Kasimir system.
The main tasks of the KASIMIR system are decision support and knowledge management for the treatment of cancer. The system is developed within a multidisciplinary research project in which participate researchers from different community (computer science, ergonomics, and oncology). For a given cancer localization, a treatment is based on a protocol similar to a medical guideline. For most of the cases (about 70%), a straightforward application of the protocol

is sufficient and provides a solution, i.e. a treatment, that can be directly reused. A case out of the 30% remaining cases is said to be *out-of-the-protocol*, i.e. either the protocol does not provide a treatment for this medical case, or the proposed solution raises some difficulties, e.g. contraindication, treatment impossibility, etc. For such an out-of-the-protocol case, oncologists try to *adapt* the protocol. In turn, these adaptations can be used to propose *evolutions* of the protocol based on a confrontation with actual cases. The idea is then to make suggestions for protocol evolutions based on frequently performed adaptations.

In knowledge-intensive CBR, the reuse of cases is generally based on adaptation, the goal of which is to solve the target problem by adapting the solution of a source case. The adaptation process is based on adaptation knowledge that –for the main part– is domain-dependent, and thus needs to be acquired for a new application of CBR. Adaptation knowledge plays a key issue in applications, e.g. in knowledge-intensive case-based reasoning systems [1].

In parallel, the Semantic Web technology relies on the availability of large amount of knowledge in various forms [16,41]. The acquisition of ontologies is one of the important issues that is widely explored in the Semantic Web community. Moreover, the acquisition of decision and adaptation knowledge for the Semantic Web has not been so deeply explored, though this kind of knowledge can be useful in numerous situations. For example, given a decision protocol and an adaptation knowledge base, the KASIMIR system can be used to apply and/or to adapt the protocol to specific medical situations.

Semi-automatic acquisition of adaptation knowledge.
The goal of *adaptation knowledge acquisition* (AKA) is to mine a case base, to extract adaptation knowledge units, and to make these units operational. Until now, the research work on CBR in the Orpailleur team has mainly focused on the design of algorithms and knowledge representation formalisms for implementing the adaptation process in a CBR system. A next step is to investigate the AKA process, a research topic that has still not received so much in the CBR community. A parallel research topic is to apply AKA to the extraction of decision knowledge units. Indeed, adaptation knowledge is closely related with decision theory, e.g. the Wald pessimistic criterion is frequently applied when pieces of information about a patient are missing.

Accordingly, the objective of the research work on AKA is to study how KDD techniques can be used for feeding a knowledge server embedded in a semantic portal –as the KASIMIR semantic portal [13]– and thus to instantiate the KDDK process. In the KASIMIR semantic portal, OWL-based formalisms for representing medical ontologies, decision protocols (the case base), and adaptation knowledge, are designed. Web services associated to the CBR process are developed. Several protocols are implemented, with a few of them including adaptation knowledge.

Practically, AKA can be considered from two main points of view. AKA from experts is based on 'manual" analysis of documents related to current problems. The AKA from expert process leads to the elaboration of *adaptation rules*, depending on formal parameters and associated with explanations. The adaptation rules are human-understandable –thanks to explanations– but they need

additional knowledge for instantiating the parameters and being applied (more on AKA from experts is given in [27,34]).

Semi-automatic AKA is based on the principles of KDD, and involves data preparation, data mining, and interpretation of the extracted units, under the control of an analyst. The input of the AKA process is a set of adaptations –thus elements at the knowledge level– and the output is a set of adaptation rules. Such an adaptation rule is an operational association rule, that lack explanations. Mixed AKA combines AKA from experts and semi-automatic AKA for supplying operational and human-understandable adaptation knowledge.

In the current experiments within the KASIMIR system, semi-automatic AKA is based on frequent itemset search. A system for AKA, named CABAMAKA–case base mining for AKA, is currently under development within the KASIMIR system and relies on semi-automatic AKA [12]. The CABAMAKA system analyzes a simple representation of the variations $\Delta\mathtt{u}$ between units of knowledge $\mathtt{u}_1$ and $\mathtt{u}_2$, where $\Delta\mathtt{u}$ encodes the substitutions transforming $\mathtt{u}_1$ into $\mathtt{u}_2$. The variations are represented in an expressive DL-based formalism, allowing a high-level expression of the extracted adaptation rules.

Beyond CBR, such a research work can be useful for ontology alignment: an alignment expresses a correspondence between the elements of two ontologies, but it could also express the variations between corresponding elements, within a rich representation formalism for the variations.

4 Towards an Integrated KDDK System

From a global point of view, the research objectives for KDDK can be summarized as follows:

- A methodology for a "knowledge discovery from complex data guided by domain knowledge process" (KDDK), i.e. a process leading from complex data units to complex knowledge units taking advantage of domain knowledge, at each step of the knowledge discovery process.
- A combination of symbolic and numerical data mining methods for setting up a complete and hybrid mining methodology to be applied on various types of data.
- An implementation of the "knowledge discovery from complex data guided by domain knowledge process" within an operational system, to be used on a large set of data types, e.g. textual documents, genomic data, spatio-temporal data, graphs, and even on sets of knowledge units (a kind of meta-knowledge mining), i.e. mining a knowledge base instead of a database.
- Accordingly, the design of a KDDK system, based on the above principles, and involved in application domains such as astronomy, agronomy, biology, chemistry, medicine, for decision support and problem-solving.

From a middle-term perspective, a system for KDDK can be considered as a "decentralized system" the architecture of which is described hereafter.

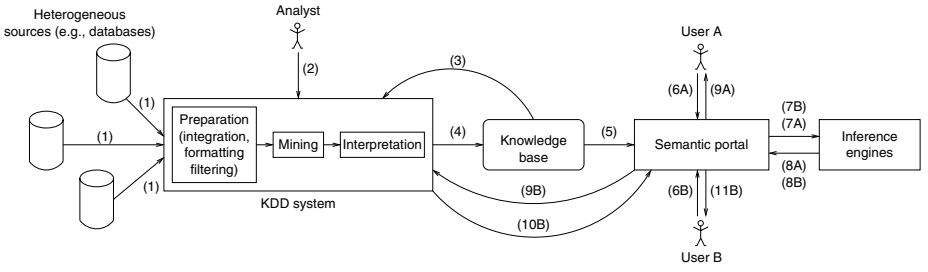

Fig. 3. An architecture for a system aimed at "knowledge discovery (from complex data) guided by domain knowledge process (KDDK)". The classical KDD process can be read from left to right, while, by contrast, the KDDK system can be read from right to left.

- One or several ontologies (knowledge bases) include knowledge from different domains with different points of view, and as well, a case base. A set of services are related through a semantic portal, for knowledge editing, navigating, and visualizing the ontologies.
- An inference engine provides, in association with the knowledge bases, a collection of inference rules for problem-solving purposes, among which subsumption, classification (lattice-based classification, clustering), case-based reasoning. Reasoning services are present for handling concrete datatypes such as strings or numbers (and possibly, for controlling procedural or functional reasoning modes if-needed).
- A set of heterogeneous databases holding on a domain to be mined for providing knowledge units enriching domain ontologies.
- A platform for KDDK proposes a collection of data mining modules –such as the CORON platform– and a set of services for data preparation and extracted unit interpretation.

Moreover, the system has to provide channels for allowing communications with human agents, such as experts and end-users. The resulting KDDK system architecture has to be reusable in any application domain. Accordingly, the integration of such a KDDK system in the framework of the semantic Web can be seen as follows. The data sources, i.e. databases, sets of documents, are explored, navigated, and queried, under the supervision of an analyst, thanks to a KDDK process guided by knowledge bases of the domain. The data are prepared and manipulated by the KDDK process, while the knowledge units are validated by the analyst, and then manipulated by the inference engine.

The figure 3 presents the architecture proposal for a KDDK system, in which different scenarios can be made operational. Heterogeneous sources (e.g. databases) feed the KDD system (1), under the supervision of an analyst (2), using available domain knowledge (3). The KDD system returns new knowledge units for extending and enriching a knowledge base (4), that may be queried through a semantic portal (5) by distant geographically distributed users (users A and B). The users A and B query the portal (6A, 6B), that in turn may use the services of a knowledge base and the associated inference engine (7A, 7B). When the

available knowledge provides, with the help of the inference engine, an answer to the request (8A), this answer is transmitted to the user (9A). Otherwise (8B), the request is transferred in a filtering module used by the KDD system (9B) for mining the available data, trying to extract information related to the request. The resulting extracted knowledge units relying on this filter (10B) may provide an answer to the user (11B).

5 Conclusion

In this paper, we have presented the research work carried out in the Orpailleur team at LORIA. Multiple and combined aspects of knowledge discovery and knowledge processing have been introduced and discussed: symbolic KDD methods such as lattice-based classification itemset search, and association rule extraction, and numeric methods such as HMM2. Next, the KDD process has been considered from a knowledge representation perspective, explaining how and why the KDD process may take advantage of domain knowledge embedded in ontologies relative to the domain of data. This perspective leads to the idea of KDDK, for knowledge discovery (from complex data) guided by domain knowledge. The KDDK process is based on classification tasks, for modeling, representing, reasoning, and discovering. Various instantiations of the KDDK process have been described, among which the mining of molecular graphs –for knowledge discovery in chemical reaction databases–, text mining and Semantic Web for designing and enlarging ontologies from documents, knowledge discovery in life sciences, and hybrid knowledge discovery, combining numerical and symbolic methods for data mining. An original experiment has also been introduced and discussed: meta-knowledge mining, or mining a knowledge base instead of a database. This research work has been carried out for the need of adaptation knowledge acquisition (AKA), that is a promising research domain, and that can be reused for mining various kind of strategical knowledge units, e.g. decision knowledge units. At the end of the paper, an architecture of an integrated KDDK system has been proposed and discussed.

References

1. Aamodt, A.: Knowledge-Intensive Case-Based Reasoning and Sustained Learning. In: Aiello, L.C. (ed.) Proc. of the 9th European Conference on Artificial Intelligence (ECAI 1990) (1990)
2. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P. (eds.): The Description Logic Handbook. Cambridge University Press, Cambridge (2003)
3. Barbut, M., Monjardet, B.: Ordre et classification – Algèbre et combinatoire (2 tomes). Hachette, Paris (1970)
4. Bendaoud, R., Rouane Hacene, M., Toussaint, Y., Delecroix, B., Napoli, A.: Text-based ontology construction using relational concept analysis. In: Flouris, G., d'Aquin, M. (eds.) Proceedings of the International Workshop on Ontology Dynamics, Innsbruck (Austria), pp. 55–68 (2007)

5. Napoli, A., Berasaluce, S., Laurenço, C., Niel, G.: An Experiment on Knowledge Discovery in Chemical Databases. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS (LNAI), vol. 3202, pp. 39–51. Springer, Heidelberg (2004)
6. Stumme, G., Berendt, B., Hotho, A.: Towards Semantic Web Mining. In: Horrocks, I., Hendler, J. (eds.) ISWC 2002. LNCS, vol. 2342, Springer, Heidelberg (2002)
7. Brachman, R.J., Selfridge, P.G., Terveen, L.G., Altman, B., Borgida, A., Halper, F., Kirk, T., Lazar, A., McGuinness, D.L., Resnick, L.A.: Knowledge representation support for data archaeology. In: Proceedings of the 1st International Conference on Information and Knowledge Management (CKIM 1992), Baltimore, pp. 457–464 (1992)
8. Carpineto, C., Romano, G.: Concept Data Analysis: Theory and Applications. John Wiley & Sons, Chichester (2004)
9. Cherfi, H., Napoli, A., Toussaint, Y.: Towards a text mining methodology using association rules extraction. Soft Computing 10(5), 431–441 (2006)
10. Cimiano, P., Hotho, A., Staab, S.: Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. Journal of Artificial Intelligence Research 24, 305–339 (2005)
11. Stumme, G., Hotho, A., Tane, J., Cimiano, P.: Conceptual Knowledge Processing with Formal Concept Analysis and Ontologies. In: Eklund, P.W. (ed.) ICFCA 2004. LNCS (LNAI), vol. 2961, pp. 189–207. Springer, Heidelberg (2004)
12. d'Aquin, M., Badra, F., Lafrogne, S., Lieber, J., Napoli, A., Szathmary, L.: Case base mining for adaptation knowledge acquisition. In: Veloso, M.M. (ed.) IJCAI 2007, Hyderabad, India, pp. 750–755. Morgan Kaufman, San Francisco (2007)
13. d'Aquin, M., Bouthier, C., Brachais, S., Lieber, J., Napoli, A.: Knowledge Edition and Maintenance Tools for a Semantic Portal in Oncology. International Journal on Human–Computer Studies 62(5), 619–638 (2005)
14. Dunham, M.H.: Data Mining – Introductory and Advanced Topics. Prentice Hall, Upper Saddle River (2003)
15. Eng, C., Thibessard, A., Hergalant, S., Mari, J.-F., Leblond, P.: Data mining using hidden markov models (HMM2) to detect heterogeneities into bacteria genomes. In: Journées Ouvertes Biologie, Informatique et Mathématiques – JOBIM 2005, Lyon, France (2005)
16. Fensel, D., Hendler, J., Lieberman, H., Wahlster, W. (eds.): Spinning the Semantic Web. The MIT Press, Cambridge, Massachusetts (2003)
17. Fuchs, B., Lieber, J., Mille, A., Napoli, A.: An Algorithm for Adaptation in Case-based Reasoning. In: Horn, W. (ed.) Proceedings of the 14th European Conference on Artificial Intelligence (ECAI-2000), Berlin, pp. 45–49. IOS Press, Amsterdam (2000)
18. Ganter, B., Wille, R.: Formal Concept Analysis. Springer, Berlin (1999)
19. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco (2001)
20. Hand, D., Mannila, H., Smyth, P.: Principles of Data Mining. The MIT Press, Cambridge (2001)
21. Janetzko, D., Cherfi, H., Kennke, R., Napoli, A., Toussaint, Y.: Knowledge-based selection of association rules for text mining. In: de Màntaras, R.L., Saitta, L. (eds.) 16h European Conference on Artificial Intelligence – ECAI 2004, Valencia, Spain, pp. 485–489 (2004)
22. Jay, N., Kohler, F., Napoli, A.: Using formal concept analysis for mining and interpreting patient flows within a healthcare network. In: Ben Yahia, S., Mephu Nguifo, E., Behlohlavek, R. (eds.) CLA 2006. LNCS (LNAI), vol. 4923, pp. 263–268. Springer, Heidelberg (2008) (this volume)

23. Kuznetsov, S.O.: Machine Learning and Formal Concept Analysis. In: Eklund, P.W. (ed.) ICFCA 2004. LNCS (LNAI), vol. 2961, pp. 287–312. Springer, Heidelberg (2004)
24. Kuznetsov, S.O., Obiedkov, S.A.: Comparing performance of algorithms for generating concept lattices. Journal of Theoretical Artificial Intelligence 14(2/3), 189–216 (2002)
25. Le Ber, F., Benoit, M., Schott, C., Mari, J.-F., Mignolet, C.: Studying crop sequences with CarrotAge, a HMM-based data mining software. Ecological Modelling 191(1), 170–185 (2006)
26. Le Ber, F., Napoli, A.: Design and comparison of lattices of topological relations for spatial representation and reasoning. Journal of Experimental & Theoretical Artificial Intelligence 15(3), 331–371 (2003)
27. Lieber, J., d'Aquin, M., Bey, P., Napoli, A., Rios, M., Sauvagnac, C.: Adaptation knowledge acquisition, a study for breast cancer treatment. In: Dojat, M., Keravnou, E.T., Barahona, P. (eds.) AIME 2003. LNCS (LNAI), vol. 2780, pp. 304–313. Springer, Heidelberg (2003)
28. Liu, H., Lu, H., Feng, L., Hussain, F.: Efficient Search of Reliable Exceptions. In: Zhong, N., Zhou, L. (eds.) PAKDD 1999. LNCS (LNAI), vol. 1574, pp. 194–204. Springer, Heidelberg (1999)
29. Mari, J.-F., Haton, J.-P., Kriouile, A.: Automatic Word Recognition Based on Second-Order Hidden Markov Models. IEEE Transactions on Speech and Audio Processing 5, 22–25 (1997)
30. Mari, J.-F., Le Ber, F.: Temporal and spatial data mining with second-order hidden models. Soft Computing 10(5), 406–414 (2006)
31. Maumus, S., Napoli, A., Szathmary, L., Visvikis-Siest, S.: Fouille de données biomédicales complexes: extraction de règles et de profils génétiques dans le cadre de l'étude du syndrome métabolique. In: Journées Ouvertes Biologie Informatique Mathématiques – JOBIM 2005,, pp. 169–173 (2005)
32. Napoli, A., Messai, N., Devignes, M.-D., Smaïl-Tabbone, M.: Querying a Bioinformatic Data Sources Registry with Concept Lattices. In: Dau, F., Mugnier, M.-L., Stumme, G. (eds.) ICCS 2005. LNCS (LNAI), vol. 3596, pp. 323–336. Springer, Heidelberg (2005)
33. Messai, N., Devignes, M.-D., Napoli, A., Smaïl-Tabbone, M.: Br-explorer: An fca-based algorithm for information retrieval. In: Ben Yahia, S., Mephu-Nguifo, E. (eds.) Fourth International Conference on Concept Lattices and their Applications (CLA-2006), Hammamet, Tunisia (2006)
34. Mollo, V.: Usage des ressources, adaptation des savoirs et gestion de l'autonomie dans la décision thérapeutique. Thèse d'Université, Conservatoire National des Arts et Métiers (2004)
35. Napoli, A.: A smooth introduction to symbolic methods for knowledge discovery. In: Cohen, H., Lefebvre, C. (eds.) Handbook of Categorization in Cognitive Science, pp. 913–933. Elsevier, Amsterdam (2005)
36. Napoli, A., Le Ber, F.: The Galois lattice as a hierarchical structure for topological relations. Annals of Mathematics and Artificial Intelligence 49(1–4), 171–190 (2007); Special volume on Knowledge discovery and discrete mathematics and a tribute to the memory of Peter L. Hammer
37. Padmanabhan, B., Tuzhilin, A.: On characterization and discovery of minimal unexpected patterns in rule discovery. IEEE Transactions on Knowledge and Data Engineering 18(2), 202–216 (2006)

38. Pennerath, F., Napoli, A.: La fouille de graphes dans les bases de données réactionnelles au service de la synthèse en chimie organique. In: Ritschard, G., Djeraba, C. (eds.) Extraction et gestion des connaissances (EGC 2006), Lille, pp. 517–528 (2006) RNTI-E-6, Cépaduès-Éditions Toulouse
39. Quan, T.T., Hui, S.C., Fong, A.C.M., Cao, T.H.: Automatic generation of ontology for scholarly semantic web. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 726–740. Springer, Heidelberg (2004)
40. Smaïl-Tabbone, M., Osman, S., Messai, N., Napoli, A., Devignes, M.-D.: Bioregistry: A structured metadata repository for bioinformatic databases. In: R. Berthold, M., Glen, R.C., Diederichs, K., Kohlbacher, O., Fischer, I. (eds.) CompLife 2005. LNCS (LNBI), vol. 3695, pp. 46–56. Springer, Heidelberg (2005)
41. Staab, S., Studer, R. (eds.): Handbook on Ontologies. Springer, Berlin (2004)
42. Stumme, G.: Formal concept analysis on its way from mathematics to computer science. In: Priss, U., Corbett, D.R., Angelova, G. (eds.) ICCS 2002. LNCS (LNAI), vol. 2393, pp. 2–19. Springer, Heidelberg (2002)
43. Suzuki, E.: Undirected Discovery of Interesting Exception Rules. International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI) 16(8), 1065–1086 (2002)
44. Szathmary, L.: Symbolic Data Mining Methods with the Coron Platform.In: Thèse d'informatique, Université Henri Poincaré – Nancy 1, France (2006)
45. Szathmary, L., Maumus, S., Petronin, P., Toussaint, Y., Napoli, A.: Vers l'extraction de motifs rares. In: Ritschard, G., Djeraba, C. (eds.) Extraction et gestion des connaissances (EGC 2006), Lille, pp. 499–510 (2006) RNTI-E-6, Cépaduès-Éditions Toulouse
46. Szathmary, L., Napoli, A.: Coron: A framework for levelwise itemset mining algorithms. In: Ganter, B., Godin, R. (eds.) ICFCA 2005. LNCS (LNAI), vol. 3403, pp. 110–113. Springer, Heidelberg (2005)
47. Szathmary, L., Napoli, A., Kuznetsov, S.O.: Zart: A multifunctional itemset mining algorithm. In: Diatta, J., Eklund, P., Liquière, M. (eds.) Proceedings of the Fifth International Conference on Concept Lattices and their Applications, Montpellier, France, pp. 26–37 (2007)
48. Szathmary, L., Napoli, A., Valtchev, P.: Towards rare itemset mining. In: Proceedings of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Patras, Greece, IEEE Computer Society Press, Los Alamitos (2007)
49. Ténier, S., Toussaint, Y., Napoli, A., Polanco, X.: Instantiation of relations for semantic annotation. In: The 2006 IEEE/WIC/ACM International Conference on Web Intelligence - WI 2006, Hong Kong, pp. 463–472. IEEE Computer Society Press, Los Alamitos (2006)
50. Ténier, S., Napoli, A., Polanco, X., Toussaint, Y.: Semantic annotation of webpages. In: Handschuh, S. (ed.) ISWC 2005. LNCS, vol. 3729, Springer, Heidelberg (2005)
51. Valtchev, P., Missaoui, R., Godin, R.: Formal concept analysis for knowledge discovery and data mining: The new challenges. In: Eklund, P.W. (ed.) ICFCA 2004. LNCS (LNAI), vol. 2961, pp. 352–371. Springer, Heidelberg (2004)
52. Weiss, G.M.: Mining with rarity: a unifying framework. SIGKDD Exploration Newsletter 6(1), 7–19 (2004)
53. Wille, R.: Methods of conceptual knowledge processing. In: Missaoui, R., Schmidt, J. (eds.) Formal Concept Analysis. LNCS (LNAI), vol. 3874, pp. 1–29. Springer, Heidelberg (2006)
54. Wu, X., Zhang, C., Zhang, S.: Efficient mining of both positive and negative association rules. ACM Transactions on Information Systems 22(3), 381–405 (2004)